VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering

J Component report

Programme : M Tech Integrated CSE
Specialization in Business Analytics.

Course Title : Exploratory Data Analysis

Course Code : CSE3039

Slot : F1

Title: "IMPACT OF DRUG ABUSE ON MODERN SOCIETY"

Team Members: SANTHAN DEEP G – 20MIA1010
SRIHARISH R – 20MIA1027
SANJAY M – 20MIA1031
JAIGANESH S – 20MIA1055

Faculty: Sweetlin Hemalatha C **Sign:**

Date: 28-04-2022

Type *Markdown* and LaTeX: α^2

In [2]:
```
from IPython.display import Image
Image("C:\\Users\\mural\\Downloads\\eda.JPEG")
```

Out[2]:

ABSTRACT

To study the impact of drug abuse that changes modern society. Addiction to drugs is a serious and often purposeful misuse of the substance, addiction occurs when a person consumes too much of a substance or drugs. The drug addiction is increasing among the people now -a- days. Most of the people in young age they are addicted to these drugs. In large number of people was found abusing drugs. We discovered a number of characteristics that contribute to drug usage among medical students, including despair, anxiety, peer pressure, mental illness, and personality disorders. Drugs are chemical substances that can change how your body and mind work. Drug addiction is a chronic brain disease. It causes a person to take drugs repeatedly, despite the harm they cause. Most of the surveys have indicated high rate of illicit and prescription drugs misuse among college students. Drug misuse, which is classified as a personality disorder, is also viewed as a global epidemic, with evolutionary genetic, pharmacological, and environmental variables impacting and controlling people's behavior. Worldwide, consumption has been at a high. Males are more likely than girls to abuse drugs, according to the study. Most of the private sector has a higher rate of drug usage.

INTRODUCTION

Substance abuse is described as a combination of different characteristics of substance use that causes to medically significant impairment or distress, as well as tolerance and symptoms of withdrawal. Emotion and psychoactive drugs aren't the only drugs that people abuse. When activity is used incorrectly for example, steroids for sports improving performance, it is considered substance abuse. Drug misuse, which is classified as a personality disorder, is also viewed as a global epidemic, with evolutionary genetic, physiologic, and environmental factors influencing and determining human behaviour. Males are more likely than girls to abuse drugs, according to the study. Most of the people with family pressures or studies or any other pressures makes them to addict these types of drugs. According to the National Institute on Drug Abuse (NIDA), 70% of high school students will have tried alcohol, 50% will have abused an illicit drug, 40% will have smoked a cigarette, and 20% will have used a prescription medication recreationally or for non - medical purposes by the time they graduate. These drug abuse may lead to social, physical, emotional, and job-related problems, most of the people are affected by this and wasted their precious life and health.

MOTIVATION

We have chosen this topic because now-a-days most of the people are addicted to these drugs and destroying their goals and future. People's take drugs for many reasons like: peer pressure, relief of stress, increased energy, to relax, to relieve pain, to escape reality, to feel more self-esteem etc. We are going to see the data and statistical view of drug abuse among the society and help young age people and students to know about the bad effects of drug addiction. And we are going to see how these drug usage impacts on the Modern society and their future. We are going to predict if he/she consumes the illicit drugs, how their future changes.

TOOLS UTILIZED

Python is an object-oriented, open-source, adaptable and simple to learn programming language. It has a rich arrangement of libraries and tools that makes things easier for data scientists. On the other hand, visual analysis allows users to visually explore data in order to discover new insights. While visual reporting arranges data navigation around defined metrics, visual analysis allows for substantially more data interaction. We can visually filter, compare, and correlate data at the speed of thought through visual analysis.

library used

In [3]:

```
#Library
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
get_ipython().run_line_magic('matplotlib', 'inline')
import pandas as pd
pd.options.display.max_columns = 30
pd.options.display.max_rows = 30
pd.set_option("display.max_columns", None)
pd.set_option("display.max_rows",None)
import numpy as np
import seaborn as sns
sns.set(rc={'figure.figsize':(12,9)})
import os
from scipy import stats
from scipy.stats import norm
from mpl_toolkits.mplot3d import Axes3D
from sklearn.preprocessing import StandardScaler
import plotly as py
import plotly.graph_objs as go
import plotly.express as px
from plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot
init_notebook_mode(connected=True)
from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_Interactivity = 'all'
```

installed wordcloud

In [4]:
```
pip install wordcloud
```

```
Requirement already satisfied: wordcloud in c:\users\mural\anaconda3\lib\site-packages (1.8.1)
Requirement already satisfied: matplotlib in c:\users\mural\anaconda3\lib\site-packages (from wordcloud) (3.3.4)
Requirement already satisfied: numpy>=1.6.1 in c:\users\mural\anaconda3\lib\site-packages (from wordcloud) (1.20.1)
Requirement already satisfied: pillow in c:\users\mural\anaconda3\lib\site-packages (from wordcloud) (8.2.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.3 in c:\users\mural\anaconda3\lib\site-package
s (from matplotlib->wordcloud) (2.4.7)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\mural\anaconda3\lib\site-packages (from matplotlib->
wordcloud) (2.8.1)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\mural\anaconda3\lib\site-packages (from matplotlib->wor
dcloud) (1.3.1)
Requirement already satisfied: cycler>=0.10 in c:\users\mural\anaconda3\lib\site-packages (from matplotlib->wordclou
d) (0.10.0)
Requirement already satisfied: six in c:\users\mural\anaconda3\lib\site-packages (from cycler>=0.10->matplotlib->wor
dcloud) (1.15.0)
Note: you may need to restart the kernel to use updated packages.
```

ABOUT OUR DATASET

This dataset is about substance abuse (cigarettes, marijuana, cocaine, alcohol) among different age groups and states. Data was collected from individual states as part of the NSDUH study. The data ranges from 2002 to 2018. Both totals (in thousands of people) and rates (as a percentage of the population) are given. The statistics give national, state, and substate estimates of substance abuse and mental disorders. The NSDUH data can also be used to predict changes over time and establish the demand for treatment services by identifying the amount of substance use and different age group with different drugs populations. Within each state and the NSDUH introduced an independent multistage area probability sample. The data set is divided into the different states and the age groups was divided upon the age group within the different usage of the illicit drugs. We can see the clear vision of the data set in the following exploration of the dataset.

In [5]:
```
#viewing the dataset
df=pd.read_csv("C:\\Users\\mural\\Downloads\\drugs (1).csv")
df
```

Out[5]:

	State	Year	Population1217	Population.18-25	Population.26+	Totals.Alcohol.Use Disorder Past Year.12-17	Totals.Alcohol.Use Disorder Past Year.18-25	Totals.Alcohol.Use Disorder Past Year.26+	Rates.Alc Disor Ye
0	Alabama	2002	380805	499453	2812905	18	68	138	
1	Alaska	2002	69400	62791	368460	4	12	27	
2	Arizona	2002	485521	602265	3329482	36	117	258	
3	Arkansas	2002	232986	302029	1687337	14	53	101	
4	California	2002	3140739	3919577	21392421	173	581	1298	
5	Colorado	2002	385648	493921	2798960	26	102	211	
6	Connecticut	2002	295157	323120	2235763	16	61	120	
7	Delaware	2002	66477	88388	514059	4	16	31	
8	District of	2002	33102	73655	372907	1	12	31	

In [6]:
```
#Starting  five datasets

df.head(5)
```

Out[6]:

	State	Year	Population1217	Population.18-25	Population.26+	Totals.Alcohol.Use Disorder Past Year.12-17	Totals.Alcohol.Use Disorder Past Year.18-25	Totals.Alcohol.Use Disorder Past Year.26+	Rates.Alcohol.Use Disorder Past Year.12-17	R
0	Alabama	2002	380805	499453	2812905	18	68	138	0.048336	
1	Alaska	2002	69400	62791	368460	4	12	27	0.061479	
2	Arizona	2002	485521	602265	3329482	36	117	258	0.073819	
3	Arkansas	2002	232986	302029	1687337	14	53	101	0.061457	
4	California	2002	3140739	3919577	21392421	173	581	1298	0.055109	

In [7]:
```
#ending five datasets

df.tail(5)
```

Out[7]:

	State	Year	Population1217	Population.18-25	Population.26+	Totals.Alcohol.Use Disorder Past Year.12-17	Totals.Alcohol.Use Disorder Past Year.18-25	Totals.Alcohol.Use Disorder Past Year.26+	Rates.Alcohol.Us Disorder Pas Year.12-1
862	Virginia	2018	629725	869285	5581639	10	82	267	0.01519
863	Washington	2018	545968	738052	5065742	11	75	282	0.02006
864	West Virginia	2018	124659	174236	1235448	2	15	46	0.01742
865	Wisconsin	2018	442510	615930	3861670	8	74	217	0.01872
866	Wyoming	2018	44908	57395	377084	1	7	22	0.02369

In [8]:
```
print("(Rows, Columns): " + str(df.shape))
df.columns
```

```
(Rows, Columns): (867, 53)
```

Out[8]:
```
Index(['State', 'Year', 'Population1217', 'Population.18-25', 'Population.26+',
       'Totals.Alcohol.Use Disorder Past Year.12-17',
       'Totals.Alcohol.Use Disorder Past Year.18-25',
       'Totals.Alcohol.Use Disorder Past Year.26+',
       'Rates.Alcohol.Use Disorder Past Year.12-17',
       'Rates.Alcohol.Use Disorder Past Year.18-25',
       'Rates.Alcohol.Use Disorder Past Year.26+',
       'Totals.Alcohol.Use Past Month.12-17',
       'Totals.Alcohol.Use Past Month.18-25',
       'Totals.Alcohol.Use Past Month.26+',
       'Rates.Alcohol.Use Past Month.12-17',
       'Rates.Alcohol.Use Past Month.18-25',
       'Rates.Alcohol.Use Past Month.26+',
       'Totals.Tobacco.Cigarette Past Month.12-17',
       'Totals.Tobacco.Cigarette Past Month.18-25',
       'Totals.Tobacco.Cigarette Past Month.26+',
       'Rates.Tobacco.Cigarette Past Month.12-17',
       'Rates.Tobacco.Cigarette Past Month.18-25',
       'Rates.Tobacco.Cigarette Past Month.26+',
       'Totals.Illicit Drugs.Cocaine Used Past Year.12-17',
       'Totals.Illicit Drugs.Cocaine Used Past Year.18-25',
       'Totals.Illicit Drugs.Cocaine Used Past Year.26+',
       'Rates.Illicit Drugs.Cocaine Used Past Year.12-17',
       'Rates.Illicit Drugs.Cocaine Used Past Year.18-25',
       'Rates.Illicit Drugs.Cocaine Used Past Year.26+',
       'Totals.Marijuana.New Users.12-17', 'Totals.Marijuana.New Users.18-25',
       'Totals.Marijuana.New Users.26+', 'Rates.Marijuana.New Users.12-17',
       'Rates.Marijuana.New Users.18-25', 'Rates.Marijuana.New Users.26+',
       'Totals.Marijuana.Used Past Month.12-17',
       'Totals.Marijuana.Used Past Month.18-25',
       'Totals.Marijuana.Used Past Month.26+',
       'Rates.Marijuana.Used Past Month.12-17',
       'Rates.Marijuana.Used Past Month.18-25',
       'Rates.Marijuana.Used Past Month.26+',
       'Totals.Marijuana.Used Past Year.12-17',
       'Totals.Marijuana.Used Past Year.18-25',
       'Totals.Marijuana.Used Past Year.26+',
       'Rates.Marijuana.Used Past Year.12-17',
       'Rates.Marijuana.Used Past Year.18-25',
       'Rates.Marijuana.Used Past Year.26+',
       'Totals.Tobacco.Use Past Month.12-17',
       'Totals.Tobacco.Use Past Month.18-25',
       'Totals.Tobacco.Use Past Month.26+',
       'Rates.Tobacco.Use Past Month.12-17',
       'Rates.Tobacco.Use Past Month.18-25',
       'Rates.Tobacco.Use Past Month.26+'],
      dtype='object')
```

In [9]:
```
#displaying total rows and columns

df['Totals.Alcohol.Use Disorder Past Year.12-17'].unique()
```

Out[9]:
```
array([ 18,   4,  36,  14, 173,  26,  16,   1,  73,   7,  10,  66,  34,
        19,  12,  17,  24,   6,  37,  57,  32,  31,   9,  13,  87,  39,
        55,  23, 118,  11,   3,  38,  35, 204,  77,  25,  54,  30,   8,
        41,  86,   5,  20,  58,  22, 113,  33,  42,  21, 200,  29,  81,
        64,  27,  51,  85,  53, 108,  15, 182,  62,  28,  47,  40,  84,
        52, 104, 186,  69,  50,  74, 105, 179,  56,  48, 103, 166,  46,
        82,  49,  98,   2, 167,  59,  43,  75,  97, 172, 133,  80,  95,
        71,  83,  70,  61,   0], dtype=int64)
```

In [10]:
```
#replacing the values for the given data


df['Totals.Alcohol.Use Disorder Past Year.12-17'] = df['Totals.Alcohol.Use Disorder Past Year.12-17'].replace(['0-14'
df['Totals.Alcohol.Use Disorder Past Year.12-17'].unique()
```

Out[10]:
```
array([ 18,   4,  36,  14, 173,  26,  16,   1,  73,   7,  10,  66,  34,
        19,  12,  17,  24,   6,  37,  57,  32,  31,   9,  13,  87,  39,
        55,  23, 118,  11,   3,  38,  35, 204,  77,  25,  54,  30,   8,
        41,  86,   5,  20,  58,  22, 113,  33,  42,  21, 200,  29,  81,
        64,  27,  51,  85,  53, 108,  15, 182,  62,  28,  47,  40,  84,
        52, 104, 186,  69,  50,  74, 105, 179,  56,  48, 103, 166,  46,
        82,  49,  98,   2, 167,  59,  43,  75,  97, 172, 133,  80,  95,
        71,  83,  70,  61,   0], dtype=int64)
```

In [11]:
```
#total count of the given data


df['Totals.Alcohol.Use Disorder Past Year.12-17'].nunique()
```

Out[11]:
```
96
```

In [12]:
```

df['Population1217'].unique()
```

Out[12]:
```
array([ 380805,   69400,  485521,  232986, 3140739,  385648,  295157,
         66477,   33192, 1346297,  748467,  103803,  128028, 1082396,
        541577,  246347,  241178,  327727,  406965,  108861,  476696,
        508325,  895753,  446545,  257508,  491394,   81697,  152465,
        184670,  113457,  719658,  176611, 1562426,  685632,   54387,
        987986,  302673,  297076, 1028108,   85295,  345629,   69742,
        473558, 2018953,  229447,   53924,  607438,  528622,  139163,
        482686,   45377,  381563,   68492,  498693,  233184, 3209345,
        388793,  295229,   67658,   33384, 1376459,  763519,  100549,
        127839, 1089901,  546518,  243608,  237856,  336908,  403315,
        109954,  485901,  512838,  902553,  442828,  256982,  490972,
         79960,  150669,  192336,  114731,  733853,  175489, 1571709,
        701982,   52643,  983181,  296942,  297247, 1034227,   87330,
        356468,   68362,  475614, 2038642,  229590,   53561,  617003,
        527419,  138269,  478926,   43883,  380150,   68162,  510698,
        232095, 3290671,  392822,  299013,   67454,   34350, 1404055,
        775954,  100585,  127552, 1099965,  549720,  240059,  233656,
        336219,  398611,  108414,  492847,  511301,  907902,  437674,
        255659,  486907,   77617,  148189,  201110,  115243,  745809,
```

In [13]:
```
df['Population1217'].nunique()
```

Out[13]:
```
866
```

In [14]:
```
df['State'].unique()
```

Out[14]:
```
array(['Alabama', 'Alaska', 'Arizona', 'Arkansas', 'California',
       'Colorado', 'Connecticut', 'Delaware', 'District of Columbia',
       'Florida', 'Georgia', 'Hawaii', 'Idaho', 'Illinois', 'Indiana',
       'Iowa', 'Kansas', 'Kentucky', 'Louisiana', 'Maine', 'Maryland',
       'Massachusetts', 'Michigan', 'Minnesota', 'Mississippi',
       'Missouri', 'Montana', 'Nebraska', 'Nevada', 'New Hampshire',
       'New Jersey', 'New Mexico', 'New York', 'North Carolina',
       'North Dakota', 'Ohio', 'Oklahoma', 'Oregon', 'Pennsylvania',
       'Rhode Island', 'South Carolina', 'South Dakota', 'Tennessee',
       'Texas', 'Utah', 'Vermont', 'Virginia', 'Washington',
       'West Virginia', 'Wisconsin', 'Wyoming'], dtype=object)
```

In [15]:
```
df['State'].nunique()
```

Out[15]:
```
51
```

In [16]:
```
df['Year'].unique()
```

Out[16]:
```
array([2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012,
       2013, 2014, 2015, 2016, 2017, 2018], dtype=int64)
```

In [17]:
```
df['Year'].nunique()
```

Out[17]:
```
17
```

In [18]:
```
# to check the missing values


df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 867 entries, 0 to 866
Data columns (total 53 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   State                                           867 non-null    object
 1   Year                                            867 non-null    int64
 2   Population1217                                  867 non-null    int64
 3   Population.18-25                                867 non-null    int64
 4   Population.26+                                  867 non-null    int64
 5   Totals.Alcohol.Use Disorder Past Year.12-17     867 non-null    int64
 6   Totals.Alcohol.Use Disorder Past Year.18-25     867 non-null    int64
 7   Totals.Alcohol.Use Disorder Past Year.26+       867 non-null    int64
 8   Rates.Alcohol.Use Disorder Past Year.12-17      867 non-null    float64
 9   Rates.Alcohol.Use Disorder Past Year.18-25      867 non-null    float64
 10  Rates.Alcohol.Use Disorder Past Year.26+        867 non-null    float64
 11  Totals.Alcohol.Use Past Month.12-17             867 non-null    int64
 12  Totals.Alcohol.Use Past Month.18-25             867 non-null    int64
 13  Totals.Alcohol.Use Past Month.26+               867 non-null    int64
 14  Rates.Alcohol.Use Past Month.12-17              867 non-null    float64
 15  Rates.Alcohol.Use Past Month.18-25              867 non-null    float64
 16  Rates.Alcohol.Use Past Month.26+                867 non-null    float64
 17  Totals.Tobacco.Cigarette Past Month.12-17       867 non-null    int64
 18  Totals.Tobacco.Cigarette Past Month.18-25       867 non-null    int64
 19  Totals.Tobacco.Cigarette Past Month.26+         867 non-null    int64
 20  Rates.Tobacco.Cigarette Past Month.12-17        867 non-null    float64
 21  Rates.Tobacco.Cigarette Past Month.18-25        867 non-null    float64
 22  Rates.Tobacco.Cigarette Past Month.26+          867 non-null    float64
 23  Totals.Illicit Drugs.Cocaine Used Past Year.12-17  867 non-null    int64
 24  Totals.Illicit Drugs.Cocaine Used Past Year.18-25  867 non-null    int64
 25  Totals.Illicit Drugs.Cocaine Used Past Year.26+    867 non-null    int64
 26  Rates.Illicit Drugs.Cocaine Used Past Year.12-17   867 non-null    float64
 27  Rates.Illicit Drugs.Cocaine Used Past Year.18-25   867 non-null    float64
 28  Rates.Illicit Drugs.Cocaine Used Past Year.26+     867 non-null    float64
 29  Totals.Marijuana.New Users.12-17                867 non-null    int64
 30  Totals.Marijuana.New Users.18-25                867 non-null    int64
 31  Totals.Marijuana.New Users.26+                  867 non-null    int64
 32  Rates.Marijuana.New Users.12-17                 867 non-null    float64
 33  Rates.Marijuana.New Users.18-25                 867 non-null    float64
 34  Rates.Marijuana.New Users.26+                   867 non-null    float64
 35  Totals.Marijuana.Used Past Month.12-17          867 non-null    int64
 36  Totals.Marijuana.Used Past Month.18-25          867 non-null    int64
 37  Totals.Marijuana.Used Past Month.26+            867 non-null    int64
 38  Rates.Marijuana.Used Past Month.12-17           867 non-null    float64
 39  Rates.Marijuana.Used Past Month.18-25           867 non-null    float64
 40  Rates.Marijuana.Used Past Month.26+             867 non-null    float64
 41  Totals.Marijuana.Used Past Year.12-17           867 non-null    int64
 42  Totals.Marijuana.Used Past Year.18-25           867 non-null    int64
 43  Totals.Marijuana.Used Past Year.26+             867 non-null    int64
 44  Rates.Marijuana.Used Past Year.12-17            867 non-null    float64
 45  Rates.Marijuana.Used Past Year.18-25            867 non-null    float64
 46  Rates.Marijuana.Used Past Year.26+              867 non-null    float64
 47  Totals.Tobacco.Use Past Month.12-17             867 non-null    int64
 48  Totals.Tobacco.Use Past Month.18-25             867 non-null    int64
 49  Totals.Tobacco.Use Past Month.26+               867 non-null    int64
 50  Rates.Tobacco.Use Past Month.12-17              867 non-null    float64
 51  Rates.Tobacco.Use Past Month.18-25              867 non-null    float64
 52  Rates.Tobacco.Use Past Month.26+                867 non-null    float64
dtypes: float64(24), int64(28), object(1)
memory usage: 359.1+ KB
```

In [19]:
```
df.isna().sum()
```

Out[19]:
```
State                                          0
Year                                           0
Population1217                                 0
Population.18-25                               0
Population.26+                                 0
Totals.Alcohol.Use Disorder Past Year.12-17    0
Totals.Alcohol.Use Disorder Past Year.18-25    0
Totals.Alcohol.Use Disorder Past Year.26+      0
Rates.Alcohol.Use Disorder Past Year.12-17     0
Rates.Alcohol.Use Disorder Past Year.18-25     0
Rates.Alcohol.Use Disorder Past Year.26+       0
Totals.Alcohol.Use Past Month.12-17            0
Totals.Alcohol.Use Past Month.18-25            0
Totals.Alcohol.Use Past Month.26+              0
Rates.Alcohol.Use Past Month.12-17             0
Rates.Alcohol.Use Past Month.18-25             0
Rates.Alcohol.Use Past Month.26+               0
Totals.Tobacco.Cigarette Past Month.12-17      0
Totals.Tobacco.Cigarette Past Month.18-25      0
Totals.Tobacco.Cigarette Past Month.26+        0
Rates.Tobacco.Cigarette Past Month.12-17       0
Rates.Tobacco.Cigarette Past Month.18-25       0
Rates.Tobacco.Cigarette Past Month.26+         0
Totals.Illicit Drugs.Cocaine Used Past Year.12-17    0
Totals.Illicit Drugs.Cocaine Used Past Year.18-25    0
Totals.Illicit Drugs.Cocaine Used Past Year.26+      0
Rates.Illicit Drugs.Cocaine Used Past Year.12-17     0
Rates.Illicit Drugs.Cocaine Used Past Year.18-25     0
Rates.Illicit Drugs.Cocaine Used Past Year.26+       0
Totals.Marijuana.New Users.12-17               0
Totals.Marijuana.New Users.18-25               0
Totals.Marijuana.New Users.26+                 0
Rates.Marijuana.New Users.12-17                0
Rates.Marijuana.New Users.18-25                0
Rates.Marijuana.New Users.26+                  0
Totals.Marijuana.Used Past Month.12-17         0
Totals.Marijuana.Used Past Month.18-25         0
Totals.Marijuana.Used Past Month.26+           0
Rates.Marijuana.Used Past Month.12-17          0
Rates.Marijuana.Used Past Month.18-25          0
Rates.Marijuana.Used Past Month.26+            0
Totals.Marijuana.Used Past Year.12-17          0
Totals.Marijuana.Used Past Year.18-25          0
Totals.Marijuana.Used Past Year.26+            0
Rates.Marijuana.Used Past Year.12-17           0
Rates.Marijuana.Used Past Year.18-25           0
Rates.Marijuana.Used Past Year.26+             0
Totals.Tobacco.Use Past Month.12-17            0
Totals.Tobacco.Use Past Month.18-25            0
Totals.Tobacco.Use Past Month.26+              0
Rates.Tobacco.Use Past Month.12-17             0
Rates.Tobacco.Use Past Month.18-25             0
Rates.Tobacco.Use Past Month.26+               0
dtype: int64
```

In [20]:
```
#to display

df.describe()
```

Out[20]:

	Year	Population1217	Population.18-25	Population.26+	Totals.Alcohol.Use Disorder Past Year.12-17	Totals.Alcohol.Use Disorder Past Year.18-25	Totals.Alcohol.Use Disorder Past Year.26+	Rates.Alcohol.Use Disorder Past Year.12-17
count	867.000000	8.670000e+02	8.670000e+02	8.670000e+02	867.000000	867.000000	867.000000	867.000000
mean	2010.000000	4.897141e+05	6.588800e+05	3.874155e+06	19.224913	94.482122	224.147636	0.040950
std	4.901807	5.637959e+05	7.559898e+05	4.320776e+06	25.290532	108.272774	254.022992	0.018241
min	2002.000000	3.055100e+04	5.739500e+04	3.101100e+05	0.000000	6.000000	19.000000	0.012143
25%	2006.000000	1.315405e+05	1.742935e+05	1.027871e+06	5.000000	26.000000	57.500000	0.025503
50%	2010.000000	3.396850e+05	4.562400e+05	2.698757e+06	11.000000	64.000000	154.000000	0.038740
75%	2014.000000	5.410950e+05	7.468080e+05	4.509094e+06	24.000000	119.500000	271.500000	0.053900
max	2018.000000	3.293484e+06	4.469106e+06	2.591772e+07	204.000000	717.000000	1586.000000	0.112131

In [21]:
```
#Describing the column state for proving the cleaning process


df['State'].describe()
```

Out[21]:
```
count            867
unique            51
top       New Jersey
freq              17
Name: State, dtype: object
```

In [22]:
```
#Gives the total count of entries for each states

df['State'].value_counts()
```

Out[22]:
```
New Jersey              17
Alabama                 17
Vermont                 17
District of Columbia    17
Missouri                17
Wyoming                 17
Colorado                17
Florida                 17
West Virginia           17
South Dakota            17
Idaho                   17
Massachusetts           17
Wisconsin               17
Maryland                17
Mississippi             17
South Carolina          17
North Carolina          17
Minnesota               17
Oklahoma                17
Montana                 17
Nevada                  17
Virginia                17
Maine                   17
Iowa                    17
Illinois                17
Washington              17
Ohio                    17
Pennsylvania            17
Arkansas                17
Delaware                17
Connecticut             17
Oregon                  17
Hawaii                  17
North Dakota            17
Michigan                17
Louisiana               17
New Mexico              17
Kansas                  17
Alaska                  17
Rhode Island            17
New Hampshire           17
Indiana                 17
Utah                    17
New York                17
Georgia                 17
Nebraska                17
Kentucky                17
California              17
Tennessee               17
Texas                   17
Arizona                 17
Name: State, dtype: int64
```

In [23]:
```
#Arranging states and their respective Totals.Alcohol.Use Past Month.12-17 count in descending order


by_states= df.groupby('State').count()['Totals.Alcohol.Use Past Month.12-17'].reset_index().sort_values(by= 'Totals.Al
by_states.style.background_gradient(cmap='Blues')
```

Out[23]:

	State	Totals.Alcohol.Use Past Month.12-17
0	Alabama	17
38	Pennsylvania	17
28	Nevada	17
29	New Hampshire	17
30	New Jersey	17
31	New Mexico	17
32	New York	17
33	North Carolina	17
34	North Dakota	17
35	Ohio	17
36	Oklahoma	17
37	Oregon	17
39	Rhode Island	17
26	Montana	17
40	South Carolina	17
41	South Dakota	17
42	Tennessee	17
43	Texas	17
44	Utah	17
45	Vermont	17
46	Virginia	17
47	Washington	17
48	West Virginia	17
49	Wisconsin	17
27	Nebraska	17
25	Missouri	17
1	Alaska	17
12	Idaho	17
2	Arizona	17
3	Arkansas	17
4	California	17
5	Colorado	17
6	Connecticut	17
7	Delaware	17
8	District of Columbia	17
9	Florida	17
10	Georgia	17
11	Hawaii	17
13	Illinois	17
24	Mississippi	17
14	Indiana	17
15	Iowa	17
16	Kansas	17
17	Kentucky	17
18	Louisiana	17
19	Maine	17
20	Maryland	17
21	Massachusetts	17
22	Michigan	17
23	Minnesota	17

	State	Totals.Alcohol.Use Past Month.12-17
50	Wyoming	17

In [24]:
```
#Checking for duplicate data


duplicate = df.duplicated()
print(duplicate.sum())
```

```
0
```

In [25]:
```
#encoding the categorical features with LabelEncoder

from sklearn.preprocessing import LabelEncoder

stat_data = df.copy()
categorical = ['State', 'Year','Totals.Alcohol.Use Disorder Past Year.12-17','Rates.Alcohol.Use Disorder Past Year.12-
le = LabelEncoder()
for column in categorical:
    stat_data[column] = le.fit_transform(stat_data[column])
stat_data.head(10)
```

Out[25]:

	State	Year	Population1217	Population.18-25	Population.26+	Totals.Alcohol.Use Disorder Past Year.12-17	Totals.Alcohol.Use Disorder Past Year.18-25	Totals.Alcohol.Use Disorder Past Year.26+	Rates.Alcohol.Use Disorder Past Year.12-17	Rate
0	0	0	380805	499453	2812905	18	68	138	550	
1	1	0	69400	62791	368460	4	12	27	738	
2	2	0	485521	602265	3329482	36	117	258	825	
3	3	0	232986	302029	1687337	14	53	101	737	
4	4	0	3140739	3919577	21392421	90	581	1298	663	
5	5	0	385648	493921	2798960	26	102	211	790	
6	6	0	295157	323120	2235763	16	61	120	644	
7	7	0	66477	88388	514059	4	16	31	697	
8	8	0	33192	73655	372907	1	12	31	292	
9	9	0	1346297	1576278	11066322	65	266	620	645	

In [26]:
```
# mean for given data


yearwise= df[['Year', 'Totals.Illicit Drugs.Cocaine Used Past Year.12-17']].groupby('Year').sum()
yearwise.reset_index(inplace = True)
round((yearwise['Totals.Illicit Drugs.Cocaine Used Past Year.12-17'].max() - yearwise['Totals.Illicit Drugs.Cocaine Us
```

Out[26]: 354.81

In [27]:
```
#correlation
stat_data.corr()
```

Out[27]:

	State	Year	Population1217	Population.18-25	Population.26+	Totals.Alcohol.Use Disorder Past Year.12-17	Totals.Alcohol.Use Disorder Past Year.18-25	T
State	1.000000e+00	-1.550186e-16	-0.072500	-0.072373	-0.076672	-0.043444	-0.069285	
Year	-1.550186e-16	1.000000e+00	-0.002989	0.024011	0.049844	-0.342138	-0.145498	
Population1217	-7.249951e-02	-2.989049e-03	1.000000	0.996437	0.989139	0.876957	0.962375	
Population.18-25	-7.237344e-02	2.401073e-02	0.996437	1.000000	0.993519	0.859589	0.957943	
Population.26+	-7.667187e-02	4.984407e-02	0.989139	0.993519	1.000000	0.846220	0.939306	
Totals.Alcohol.Use Disorder Past Year.12-17	-4.344417e-02	-3.421385e-01	0.876957	0.859589	0.846220	1.000000	0.940835	

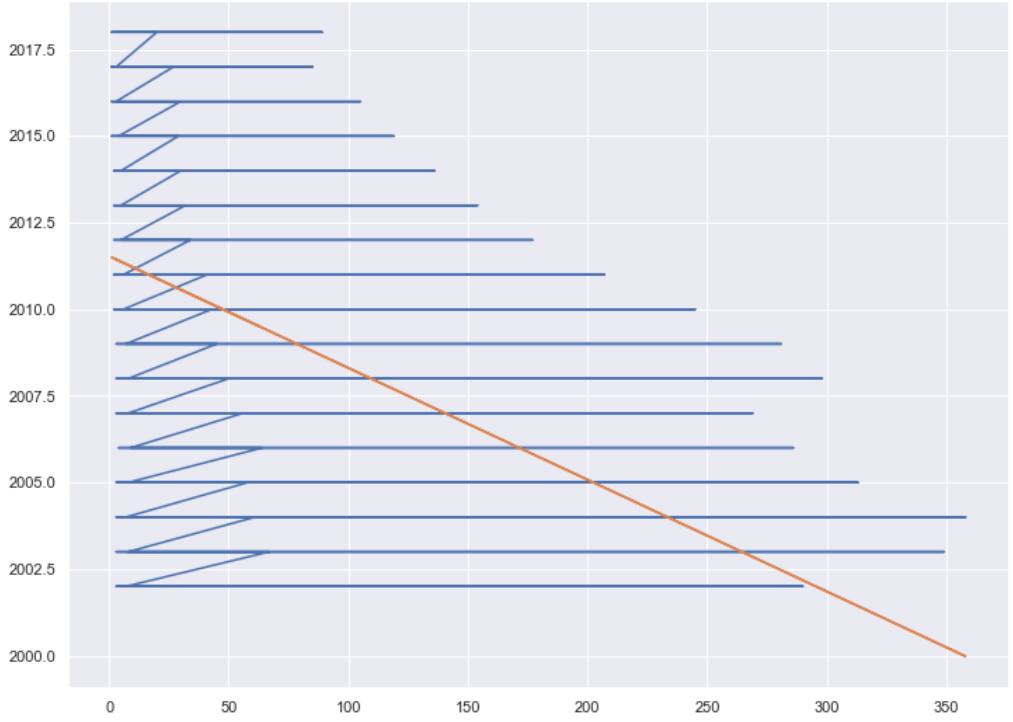
In [28]:
```
#freqency of the year


Year_column = df.loc[:,'Year']
Year = Year_column.values
Year
```

Out[28]:
```
array([2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002,
       2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002,
       2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002,
       2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002, 2002,
       2002, 2002, 2002, 2002, 2002, 2002, 2002, 2003, 2003, 2003, 2003,
       2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003,
       2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003,
       2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003,
       2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003, 2003,
       2003, 2003, 2003, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004,
       2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004,
       2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004,
       2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004,
       2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2004, 2005,
       2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005,
       2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005,
       2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005,
       2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005, 2005,
       2005, 2005, 2005, 2005, 2005, 2005, 2005, 2006, 2006, 2006, 2006,
       2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006,
       2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006,
       2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006,
       2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006, 2006,
       2006, 2006, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007,
       2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007,
       2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007,
       2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007,
       2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2008, 2008,
       2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008,
       2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008,
       2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008,
       2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008, 2008,
       2008, 2008, 2008, 2008, 2008, 2009, 2009, 2009, 2009, 2009, 2009,
       2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009,
       2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009,
       2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009,
       2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009, 2009,
       2009, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010,
       2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010,
       2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010,
       2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010,
       2010, 2010, 2010, 2010, 2010, 2010, 2010, 2010, 2011, 2011, 2011,
       2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011,
       2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011,
       2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011,
       2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011, 2011,
       2011, 2011, 2011, 2011, 2012, 2012, 2012, 2012, 2012, 2012, 2012,
       2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012,
       2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012,
       2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012,
       2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012,
       2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012, 2012,
       2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,
       2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,
       2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,
       2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013, 2013,
       2013, 2013, 2013, 2013, 2013, 2013, 2013, 2014, 2014, 2014, 2014,
       2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014,
       2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014,
       2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014,
       2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014,
       2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014,
       2014, 2014, 2014, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015,
       2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015,
       2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015,
       2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015,
       2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2015, 2016,
       2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016,
       2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016,
       2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016,
       2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016,
       2016, 2016, 2016, 2016, 2016, 2016, 2017, 2017, 2017, 2017, 2017,
       2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017,
       2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017,
       2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017,
       2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017, 2017,
       2017, 2017, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018,
       2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018,
       2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018,
       2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018,
       2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018], dtype=int64)
```

In [29]:
```
#totol column values for the given data


Total_column = df.loc[:,'Totals.Tobacco.Use Past Month.12-17']
Total = Total_column.values
Total
```

Out[29]:
```
array([ 63,  11,  73,  46, 290,  67,  45,  11,   3, 193, 118,  13,  19,
       162,  91,  40,  38,  76,  72,  15,  61,  73, 142,  79,  43, 103,
        17,  29,  28,  19,  98,  27, 200, 123,  11, 171,  53,  43, 172,
        14,  54,  16,  81, 272,  19,  10,  94,  70,  30,  84,   7,  67,
        11,  73,  41, 349,  67,  49,  10,   3, 170, 113,  12,  19, 145,
        79,  36,  41,  82,  60,  17,  59,  70, 140,  71,  39, 100,  17,
        24,  26,  18, 103,  29, 196, 121,  11, 166,  55,  44, 164,  14,
        54,  15,  77, 264,  20,   9,  83,  69,  27,  86,   8,  61,  10,
        70,  38, 358,  63,  46,  10,   3, 168, 100,  10,  17, 141,  78,
        35,  39,  71,  49,  19,  59,  71, 127,  69,  35,  94,  15,  23,
        26,  16,  99,  26, 202, 122,   9, 157,  52,  43, 160,  13,  54,
        13,  80, 243,  24,   9,  81,  61,  27,  78,   7,  58,  10,  70,
        38, 313,  55,  37,   9,   3, 162,  97,   9,  16, 141,  78,  39,
        37,  66,  50,  17,  54,  73, 121,  67,  34,  78,  14,  21,  25,
        15,  95,  26, 189, 103,   8, 154,  48,  42, 156,  12,  54,  12,
        80, 243,  23,   9,  80,  67,  25,  65,   8,  64,   9,  66,  38,
       286,  57,  37,   8,   4, 162, 113,  10,  15, 138,  79,  36,  36,
        69,  50,  14,  50,  67, 118,  63,  31,  70,  13,  19,  25,  14,
        84,  26, 163,  98,   8, 159,  49,  40, 154,  12,  51,  11,  79,
       239,  19,   8,  77,  70,  23,  72,   9,  56,   9,  60,  35, 269,
        55,  33,   9,   3, 156, 110,  10,  14, 140,  81,  31,  31,  61,
        43,  13,  44,  55, 114,  59,  29,  68,  12,  20,  25,  14,  69,
        22, 163,  99,   7, 142,  43,  43, 144,  10,  44,  10,  69, 222,
        19,   7,  72,  67,  23,  66,   8,  50,   8,  55,  35, 298,  51,
        32,   9,   3, 143,  96,   9,  14, 133,  74,  30,  30,  57,  40,
        12,  42,  52, 109,  53,  31,  70,  12,  18,  26,  13,  69,  21,
       165,  87,   7, 133,  40,  44, 136,   8,  42,  10,  62, 212,  17,
         6,  71,  62,  22,  58,   8,  45,   8,  62,  32, 281,  47,  31,
         8,   3, 139,  95,   9,  16, 115,  64,  31,  31,  50,  40,  12,
        42,  51,  98,  46,  34,  75,  12,  17,  21,  14,  68,  21, 151,
        80,   6, 130,  42,  37, 125,   9,  41,   9,  65, 210,  19,   6,
        65,  58,  23,  52,   7,  43,   7,  56,  30, 245,  46,  31,   8,
         2, 120,  93,  10,  15, 103,  56,  31,  28,  50,  40,  11,  36,
        57,  91,  44,  32,  70,  12,  16,  21,  15,  77,  20, 133,  82,
         6, 118,  41,  30, 119,   8,  42,   9,  64, 198,  18,   6,  57,
        59,  22,  46,   6,  41,   6,  43,  27, 207,  42,  25,   7,   2,
       100,  85,   8,  13,  90,  59,  25,  25,  49,  37,  10,  35,  54,
        88,  45,  32,  58,  11,  14,  20,  11,  62,  19, 117,  71,   6,
       108,  38,  26, 110,   8,  44,   8,  58, 183,  17,   6,  54,  50,
        20,  44,   6,  34,   6,  38,  25, 177,  35,  20,   6,   2, 101,
        69,   8,  11,  79,  56,  23,  21,  41,  40,   9,  30,  39,  78,
        38,  30,  54,   9,  14,  18,  10,  48,  17, 105,  61,   6,  94,
        30,  25, 102,   7,  36,   6,  52, 161,  16,   5,  53,  44,  17,
        43,   5,  32,   6,  37,  26, 154,  37,  20,   6,   2,  92,  58,
         6,  12,  69,  49,  23,  21,  37,  36,   8,  27,  32,  65,  32,
        25,  50,   7,  13,  17,   9,  50,  13,  85,  57,   5,  86,  34,
        25,  86,   5,  28,   7,  50, 137,  15,   4,  48,  41,  16,  42,
         5,  30,   6,  34,  23, 131,  29,  18,   5,   2,  72,  52,   5,
        10,  61,  49,  22,  18,  35,  29,   8,  28,  29,  60,  28,  22,
        44,   7,  11,  15,   8,  37,  10,  81,  46,   5,  72,  32,  20,
        68,   4,  26,   6,  40, 136,  11,   4,  38,  34,  16,  36,   5,
        29,   6,  24,  18, 119,  20,  15,   3,   1,  62,  50,   4,   8,
        61,  42,  16,  14,  33,  29,   8,  25,  23,  52,  24,  20,  40,
         8,   8,  12,   7,  32,   8,  69,  46,   4,  63,  23,  18,  59,
         5,  27,   5,  38, 106,  10,   4,  35,  29,  16,  28,   4,  30,
         5,  20,  18, 105,  20,  11,   3,   1,  65,  47,   3,   9,  52,
        35,  12,  13,  33,  25,   7,  20,  21,  43,  23,  19,  34,   6,
         8,  10,   6,  29,   8,  52,  46,   4,  62,  20,  16,  54,   4,
        25,   6,  36,  91,  11,   3,  32,  29,  12,  25,   3,  27,   4,
        18,  18,  83,  21,   9,   3,   1,  65,  38,   3,   8,  38,  26,
        11,  10,  31,  21,   6,  19,  24,  38,  21,  18,  28,   5,   6,
        10,   5,  25,   9,  49,  42,   3,  57,  17,  14,  45,   3,  27,
         5,  33,  85,   9,   3,  28,  31,  10,  20,   3,  20,   4,  16,
        13,  77,  18,   8,   3,   1,  48,  32,   2,   6,  34,  27,  12,
         8,  24,  21,   5,  16,  19,  34,  20,  17,  24,   6,   6,   8,
         4,  20,   8,  43,  38,   3,  49,  17,  13,  40,   2,  22,   4,
        29,  89,   8,   3,  28,  20,  10,  17,   3], dtype=int64)
```

In [30]:
```
#trend line for correlation(tobacco usage)


plt. plot(Total, Year, '-')
m, b = np. polyfit(Total, Year, 1)
x = plt. plot(Total, m*Total + b)
```



In [31]:
```
#Creating Contingency Table

contingency_table = pd.crosstab(stat_data.State, stat_data.Year)

#Significance Level 5%

alpha=0.05
```

In [32]:
```
chistat, p, dof, expected = stats.chi2_contingency(contingency_table )
```

In [33]:
```
#critical value

critical_value=stats.chi2.ppf(q=1-alpha,df=dof)
print('critical_value:',critical_value)
```

```
critical_value: 866.9114021093793
```

In [34]:
```
print('Significance level: ',alpha)
print('Degree of Freedom: ',dof)
print('chi-square statistic:',chistat)
print('critical_value:',critical_value)
print('p-value:',p)
```

```
Significance level:  0.05
Degree of Freedom:  800
chi-square statistic: 0.0
critical_value: 866.9114021093793
p-value: 1.0
```

In [35]:
```
#dependency b/w critical value and chi-square


if chistat>=critical_value:
    print("Reject H0,There is a dependency between State & Year.")
else:
    print("Retain H0,There is no relationship between State & Year.")
```

```
Retain H0,There is no relationship between State & Year.
```

In [36]:
```
if p<=alpha:
    print("Reject H0,There is a dependency between State & Year.")
else:
    print("Retain H0,There is no relationship between State & Year.")
```

```
Retain H0,There is no relationship between State & Year.
```

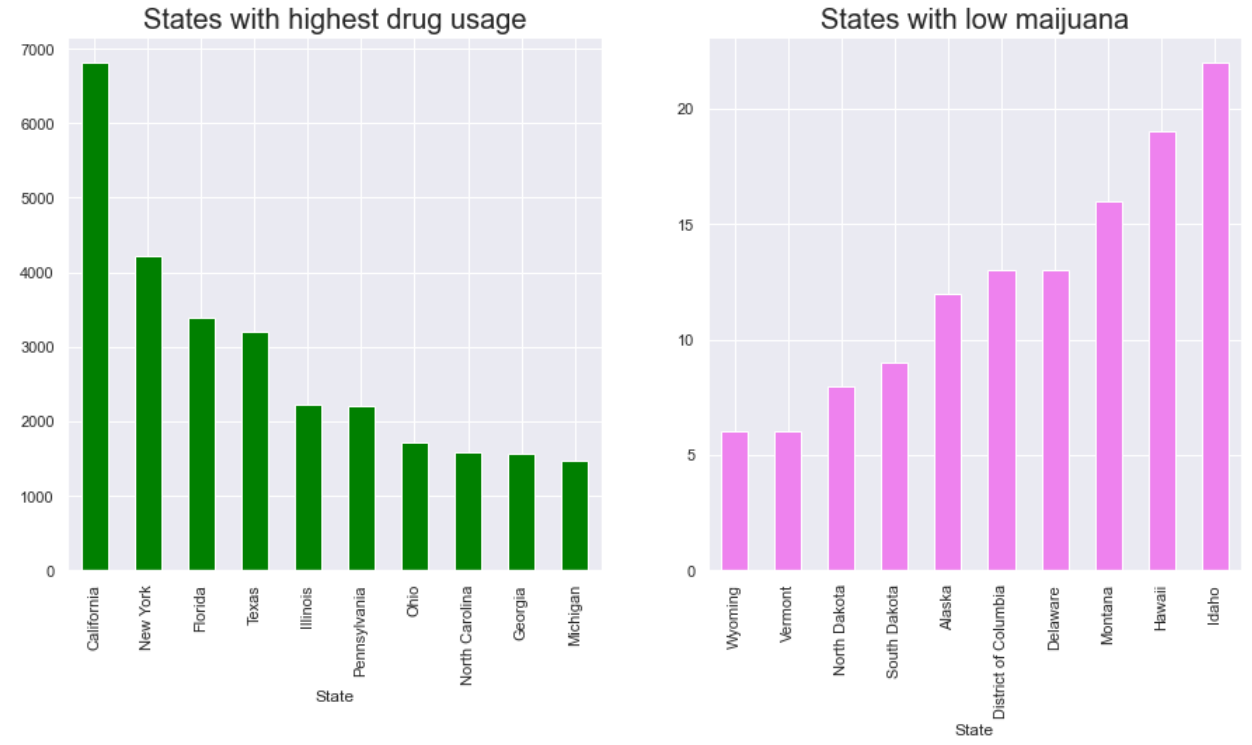
In [37]:
```
#ascending order

df.groupby("State")["Totals.Marijuana.Used Past Year.26+"].sum().sort_values(ascending = False)[:10]
```

Out[37]:
```
State
California       43018
New York         21352
Florida          18741
Texas            16590
Michigan         11779
Pennsylvania     11616
Illinois         11447
Ohio             10779
Washington        9950
Massachusetts     8885
Name: Totals.Marijuana.Used Past Year.26+, dtype: int64
```

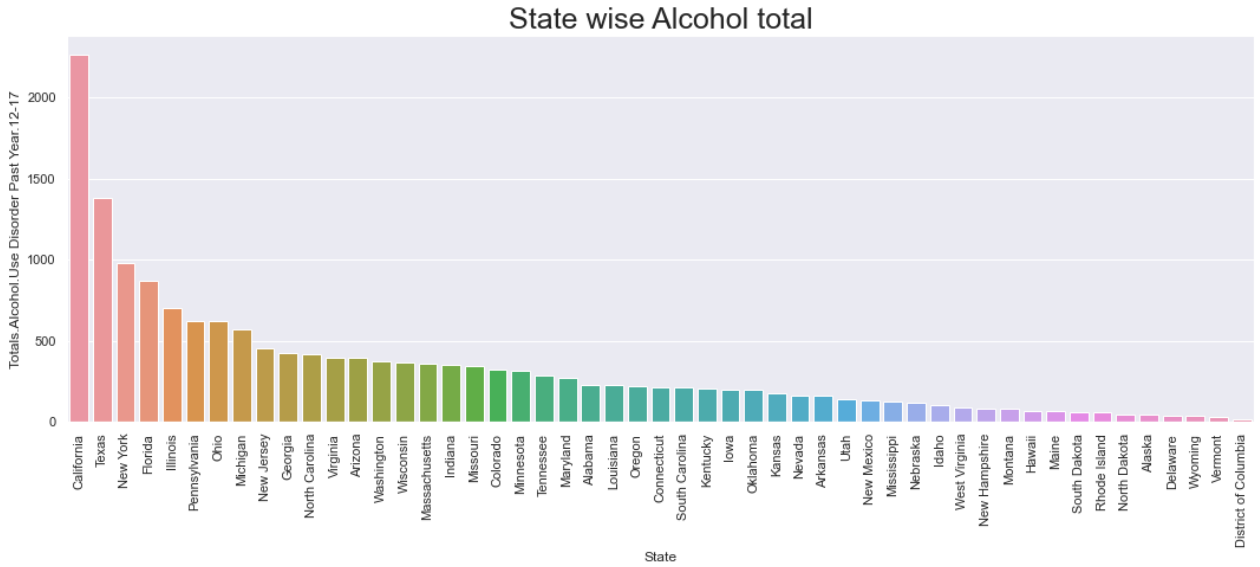
In [38]:
```
#descending order

df.groupby("State")["Totals.Marijuana.Used Past Year.26+"].sum().sort_values(ascending = True)[:10]
```

Out[38]:
```
State
North Dakota             440
Wyoming                  448
South Dakota             578
Delaware                 921
Vermont                 1008
District of Columbia    1094
Alaska                  1117
Idaho                   1285
Montana                 1295
Nebraska                1326
Name: Totals.Marijuana.Used Past Year.26+, dtype: int64
```

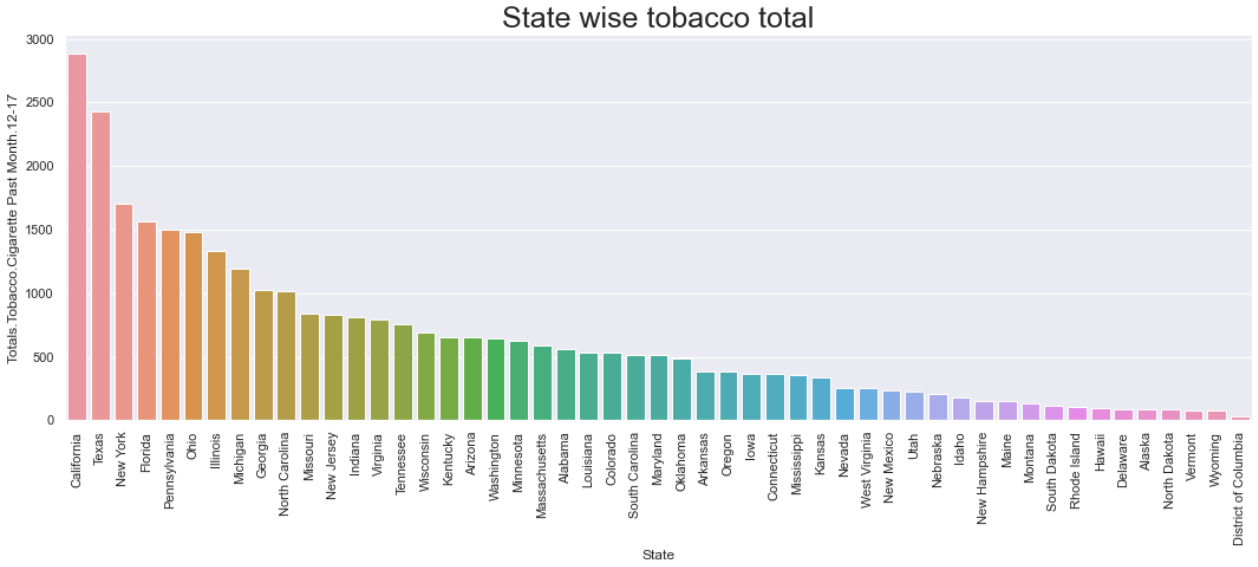
In [39]:
```
#graphical view


f, ax = plt.subplots(1,2, figsize = (15,7))
df.groupby("State")["Totals.Illicit Drugs.Cocaine Used Past Year.26+"].sum().sort_values(ascending = False)[:10].plot(
df.groupby("State")["Totals.Marijuana.New Users.26+"].sum().sort_values(ascending = True)[:10].plot(kind = "bar", colo
ax[0].set_title('States with highest drug usage', fontsize = 20)
x = ax[1].set_title('States with low maijuana', fontsize = 20)
```



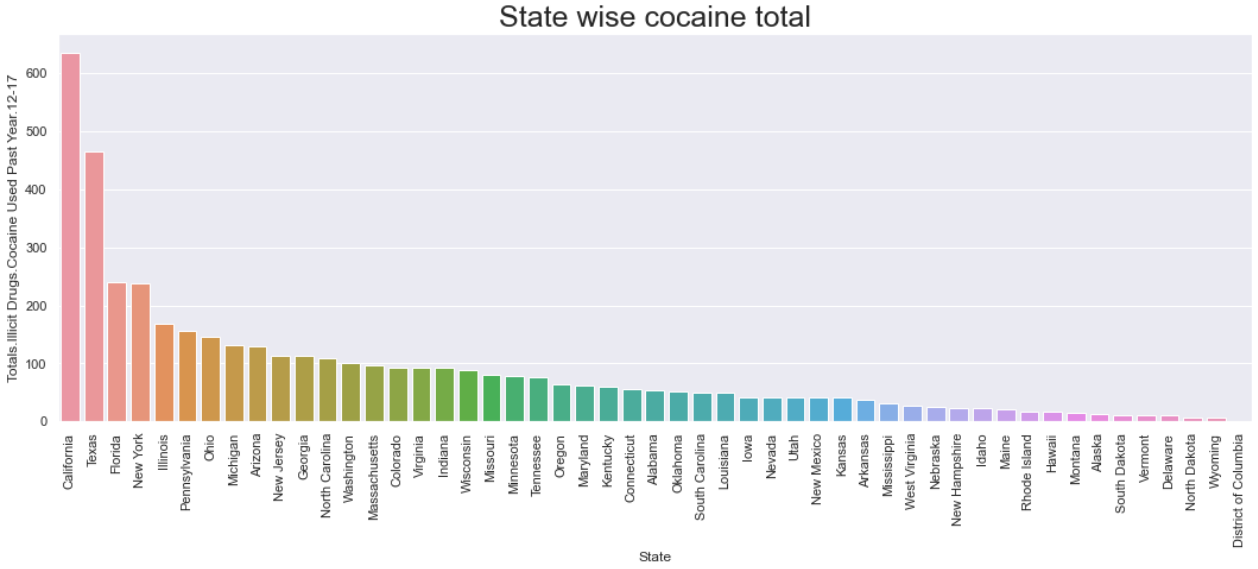
In [40]:
```
#state wise alcohol

grp = df.groupby('State')['Totals.Alcohol.Use Disorder Past Year.12-17'].sum()
total_drugs = pd.DataFrame(grp).reset_index().sort_values('Totals.Alcohol.Use Disorder Past Year.12-17',ascending=Fals

fig , ax = plt.subplots(figsize=(18,6))
g = sns.barplot(x = 'State', y = 'Totals.Alcohol.Use Disorder Past Year.12-17',data = total_drugs,ax=ax)
g.set_xticklabels(g.get_xticklabels(),rotation=90)
x = g.set_title('State wise Alcohol total', fontsize = 25)
```



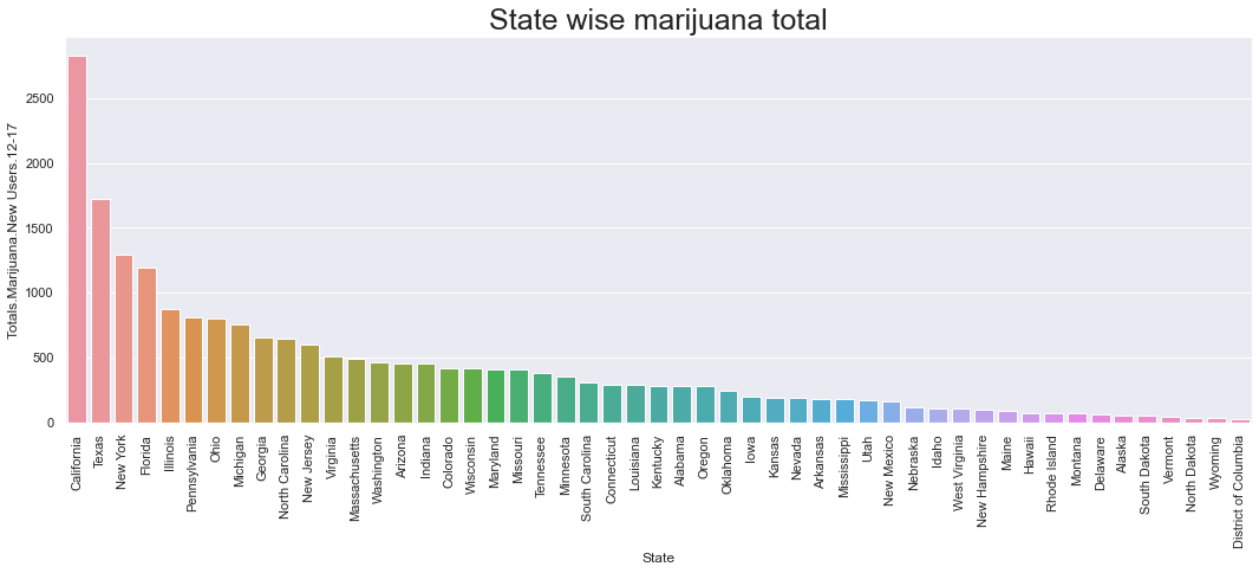
In [41]:
```
#state wise tobacco total

grp = df.groupby('State')['Totals.Tobacco.Cigarette Past Month.12-17'].sum()
total_drugs = pd.DataFrame(grp).reset_index().sort_values('Totals.Tobacco.Cigarette Past Month.12-17',ascending=False)

fig , ax = plt.subplots(figsize=(18,6))
g = sns.barplot(x = 'State', y = 'Totals.Tobacco.Cigarette Past Month.12-17',data = total_drugs,ax=ax)
g.set_xticklabels(g.get_xticklabels(),rotation=90)
x = g.set_title('State wise tobacco total', fontsize = 25)
```



In [42]:
```
#state wise cocaine total

grp = df.groupby('State')['Totals.Illicit Drugs.Cocaine Used Past Year.12-17'].sum()
total_drugs = pd.DataFrame(grp).reset_index().sort_values('Totals.Illicit Drugs.Cocaine Used Past Year.12-17',ascendir

fig , ax = plt.subplots(figsize=(18,6))
g = sns.barplot(x = 'State', y = 'Totals.Illicit Drugs.Cocaine Used Past Year.12-17',data = total_drugs,ax=ax)
g.set_xticklabels(g.get_xticklabels(),rotation=90)
x = g.set_title('State wise cocaine total', fontsize = 25)
```



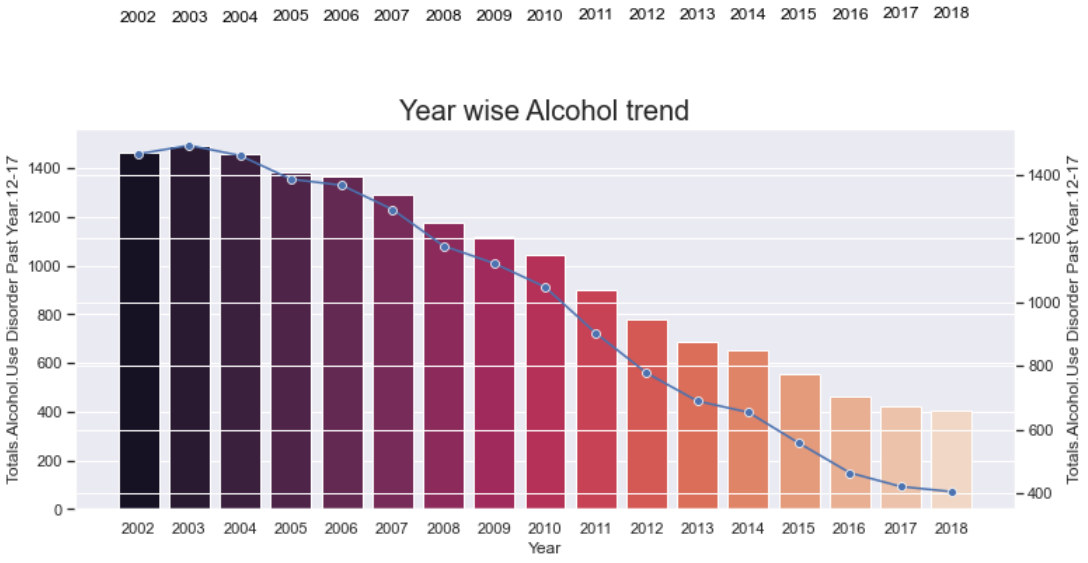
In [43]:
```
#state wise marijuana total

grp = df.groupby('State')['Totals.Marijuana.New Users.12-17'].sum()
total_drugs = pd.DataFrame(grp).reset_index().sort_values('Totals.Marijuana.New Users.12-17',ascending=False)

fig , ax = plt.subplots(figsize=(18,6))
g = sns.barplot(x = 'State', y = 'Totals.Marijuana.New Users.12-17',data = total_drugs,ax=ax)
g.set_xticklabels(g.get_xticklabels(),rotation=90)
x = g.set_title('State wise marijuana total', fontsize = 25)
```



In [44]:
```
reasons_set = df[df['Year'] == '2002']
states_set = reasons_set[['Population1217','State','Totals.Alcohol.Use Disorder Past Year.12-17']]
states = reasons_set['State'].value_counts().index
states = list(states)
for x in states:
    grp_set = states_set[states_set['State'] == x ]
    grp_set =grp_set.groupby('Population1217').sum().sort_values('Totals.Alcohol.Use Disorder Past Year.12-17', ascend
    grp_set = grp_set.head(10)
    grp_set.plot(kind = 'bar', figsize = (15,5))
    plt.title(x+ ' Alcohol disorder', fontsize = 15)
    plt.show()
```

In [45]:
```
#word visivalization

filter_state = pd.DataFrame(df.groupby(["State"])["Year"].sum()).reset_index()
from wordcloud import WordCloud
count = {}
for x in filter_state["State"].values:
    count[x]=int(filter_state[filter_state["State"]==x].Year)


wordcloud = WordCloud(width=1280,height=720,background_color='white').generate_from_frequencies(count)
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

In [46]:
```
#year wise alcohol trend

grp_yr = df.groupby('Year')['Totals.Alcohol.Use Disorder Past Year.12-17'].sum()
year = pd.DataFrame(grp_yr).reset_index().sort_values('Year',ascending=False)
fig, ax1 = plt.subplots(figsize=(12,5))
y = sns.barplot(x='Year',y='Totals.Alcohol.Use Disorder Past Year.12-17',data=year,palette = 'rocket',ax=ax1)
for index, row in year.iterrows():
    y.text(x = row.name, y = row.Year, s = str(row.Year),color='black', ha="center")
ax2 = ax1.twinx()
g=sns.lineplot(data = year['Totals.Alcohol.Use Disorder Past Year.12-17'], marker='o', sort = False, ax=ax2)
x = g.set_title('Year wise Alcohol trend', fontsize = 20)
```



In [47]:
```
#index setting

year = year.set_index('Year')
year
```

Out[47]:

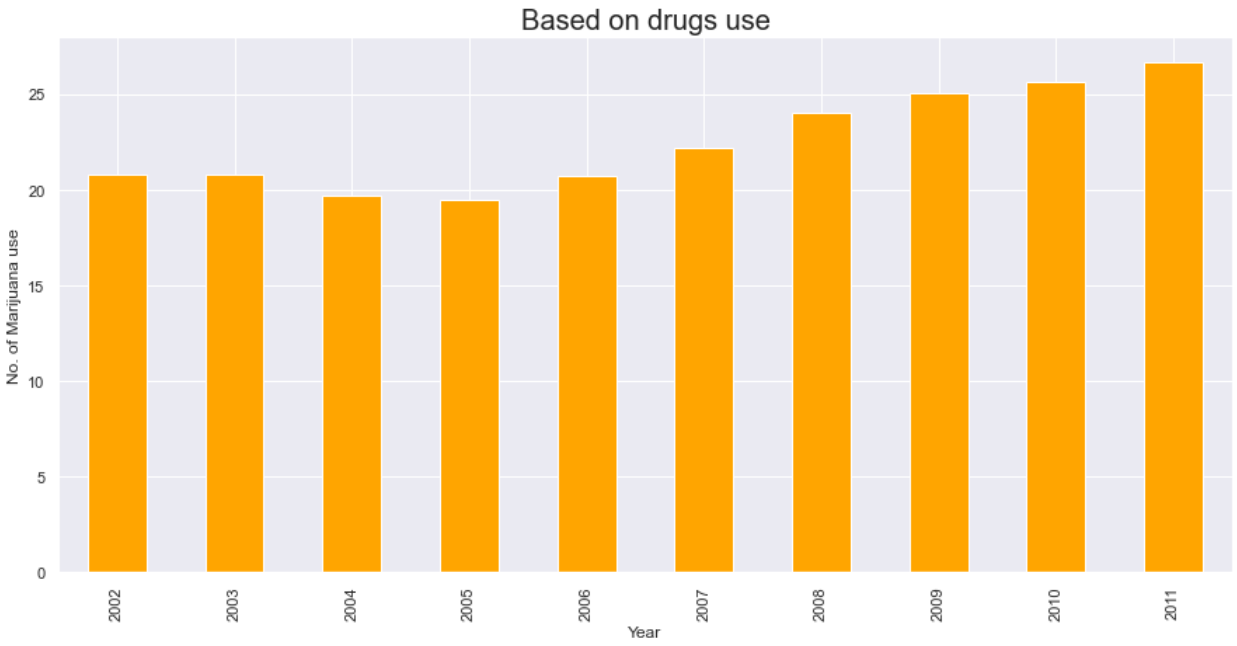
Year	Totals.Alcohol.Use Disorder Past Year.12-17
2018	405
2017	421
2016	464
2015	557
2014	654
2013	689
2012	777
2011	902
2010	1047
2009	1121
2008	1176
2007	1290
2006	1366
2005	1385
2004	1459
2003	1491
2002	1464

In [ ]:
```

```

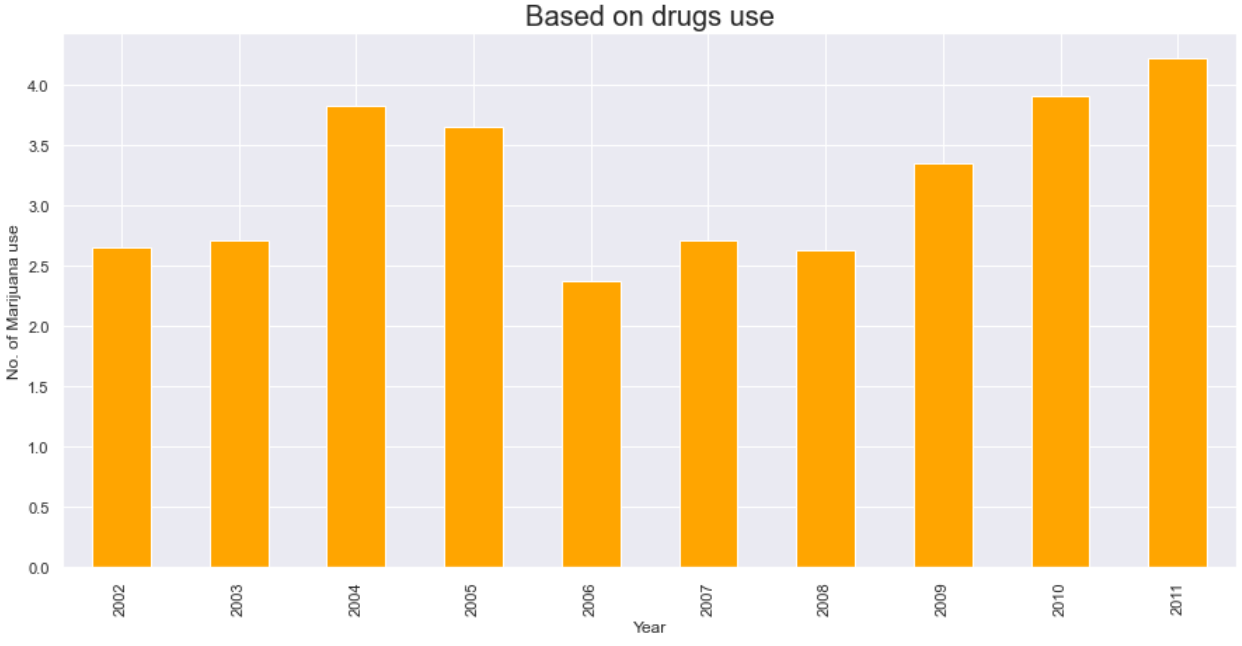
In [49]:
```
#Year wise, how the reasons are changing

year_reasons = reasons_set[['Year', 'State', 'Totals.Alcohol.Use Past Month.12-17']]
year_reasons = year_reasons.groupby(['State', 'Year']).sum().reset_index()
reasons = year_reasons['State']
reasons = reasons.value_counts()
years = year_reasons['Year'].values
years = list(years)
count = 1
reasons = list(reasons.index)
for var in reasons:
    plt.rcParams.update({'font.size': 10})
    fig = plt.figure(figsize = (10,5))
    trace1 = year_reasons[year_reasons['State'] == var]
    plt.plot( 'Year', 'Totals.Alcohol.Use Past Month.12-17', data=trace1, marker='o', markerfacecolor='blue', markersi
    plt.title(var + '--Reason Trend', fontsize = 15)
    plt.tight_layout()
    plt.show()
```

In [50]:
```
age_set = reasons_set[['State','Year','Totals.Marijuana.New Users.18-25']]
age_grp = reasons_set['Year'].value_counts().index
age_grp = list(age_grp)
for x in age_grp:
    group_set = age_set[age_set['Year'] == x ]
    group_set =group_set.groupby('State').sum().sort_values('Totals.Marijuana.New Users.18-25', ascending = False)
    group_set = group_set.head(10)
    group_set.plot(kind = 'bar', figsize = (15,5))
    plt.title('State '+x+ ' drug usage Reasons', fontsize = 15)
    plt.show()
```

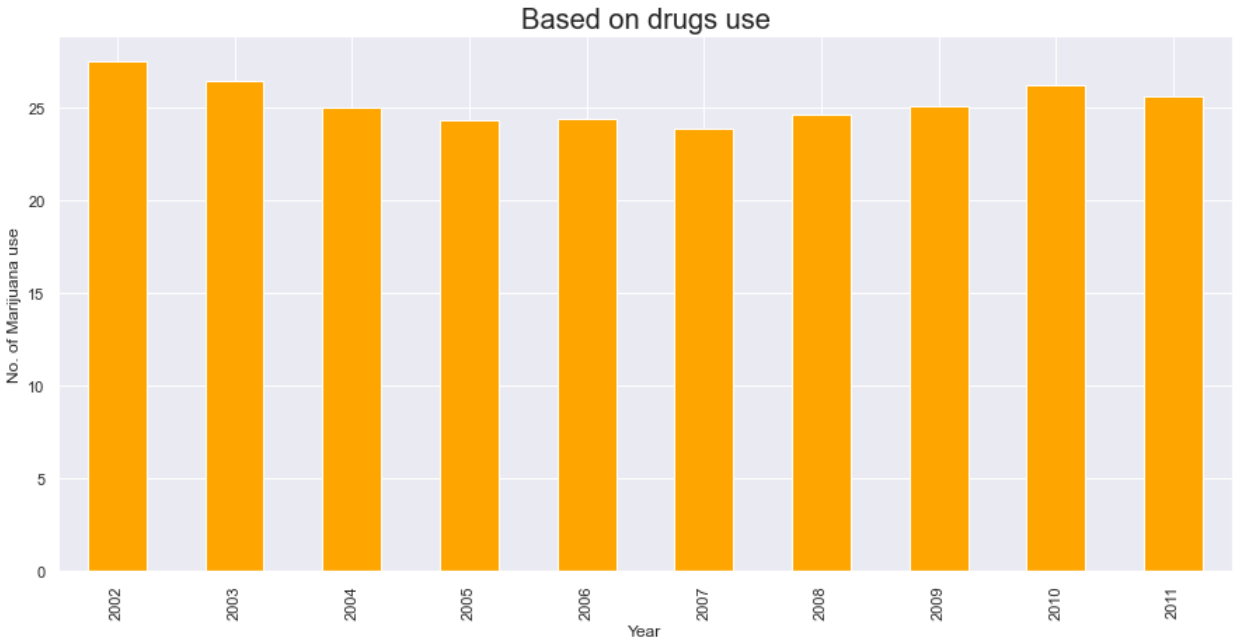
In [51]:
```
#Marjuna Drug usage by year .

df.groupby("Year")["Totals.Marijuana.New Users.18-25"].mean()[:10].plot(kind = "bar", color = "orange", figsize = (15,
plt.ylabel("No. of Marijuana use")
x = plt.title("Based on drugs use", fontsize = 20)
```



In [52]:
```
#Marjuna new users 26+ by year wise.

df.groupby("Year")["Totals.Marijuana.New Users.26+"].mean()[:10].plot(kind = "bar", color = "orange", figsize = (15,7)
plt.ylabel("No. of Marijuana use")
x = plt.title("Based on drugs use", fontsize = 20)
```



In [53]:
```
df.groupby("Year")["Totals.Marijuana.New Users.12-17"].mean()[:10].plot(kind = "bar", color = "orange", figsize = (15,
plt.ylabel("No. of Marijuana use")
x = plt.title("Based on drugs use", fontsize = 20)
```



In [54]:
```
#outlier analysis of the given data of total alcohol disorder of past year 12-17

Total_column = df.loc[:,'Totals.Alcohol.Use Disorder Past Year.12-17']
Total = Total_column.values
Total
```

Out[54]:
```
array([ 18,   4,  36,  14, 173,  26,  16,   4,   1,  73,  36,   7,  10,
        66,  34,  19,  12,  17,  24,   6,  26,  37,  57,  32,  10,  31,
         9,  14,  13,   9,  37,  14,  87,  39,   6,  57,  17,  17,  55,
         6,  17,   7,  23, 118,  11,   3,  38,  31,   9,  35,   3,  18,
         4,  37,  13, 204,  26,  19,   3,   1,  77,  32,   7,   9,  66,
        31,  18,  17,  18,  24,   7,  25,  34,  54,  30,  10,  35,   8,
        12,  13,   7,  41,  14,  86,  32,   5,  55,  20,  19,  58,   6,
        19,   7,  22, 113,  11,   3,  33,  34,   9,  42,   3,  21,   4,
        36,  12, 200,  29,  19,   3,   1,  81,  33,   5,   9,  64,  33,
        18,  18,  20,  20,   7,  25,  27,  51,  27,  11,  34,   7,  11,
        13,   7,  42,  12,  85,  38,   4,  53,  18,  19,  58,   5,  17,
         5,  23, 108,  12,   3,  31,  30,   9,  38,   3,  19,   4,  30,
        15, 182,  27,  18,   3,   1,  73,  36,   5,   9,  62,  28,  20,
        17,  20,  17,   6,  25,  28,  47,  29,  11,  31,   8,  11,  13,
         7,  40,  11,  84,  35,   4,  52,  15,  19,  53,   5,  16,   5,
        24, 104,  12,   3,  31,  31,   7,  29,   3,  19,   4,  26,  17,
       186,  28,  21,   3,   1,  69,  35,   6,   9,  64,  27,  18,  16,
        17,  17,   6,  25,  30,  50,  31,  10,  29,   6,  10,  13,   7,
        38,  11,  74,  31,   4,  55,  14,  16,  50,   5,  18,   5,  27,
       105,  11,   3,  30,  30,   6,  30,   3,  17,   3,  32,  12, 179,
        27,  18,   3,   1,  66,  32,   5,   8,  56,  28,  15,  14,  14,
        15,   5,  22,  27,  47,  29,  10,  28,   5,   9,  13,   6,  32,
        10,  74,  31,   3,  53,  14,  17,  48,   4,  16,   4,  22, 103,
        10,   3,  32,  29,   6,  30,   3,  16,   3,  29,  11, 166,  24,
        17,   3,   1,  58,  28,   5,   7,  46,  27,  16,  13,  13,  13,
         4,  17,  24,  41,  24,   9,  25,   6,   8,  12,   5,  26,  11,
        82,  25,   3,  49,  15,  18,  41,   4,  14,   4,  16,  98,   8,
         2,  31,  25,   6,  24,   3,  16,   3,  26,  11, 167,  22,  14,
         3,   1,  59,  30,   5,   7,  43,  23,  13,  12,  14,  12,   4,
        15,  25,  37,  20,   9,  22,   5,   7,  11,   5,  27,   9,  75,
        26,   2,  43,  14,  15,  40,   4,  14,   4,  18,  97,  10,   2,
        30,  22,   6,  20,   2,  14,   3,  25,   9, 172,  18,  13,   3,
         1,  55,  31,   5,   6,  41,  19,  10,  10,  12,  12,   3,  15,
        24,  33,  18,   9,  20,   4,   5,  11,   5,  33,   7,  62,  29,
         2,  36,  11,  12,  40,   3,  14,   3,  18,  85,  10,   2,  26,
        24,   5,  17,   2,  13,   2,  20,   7, 133,  15,  11,   2,   1,
        42,  25,   4,   5,  37,  19,   8,   8,  11,  12,   3,  15,  20,
        31,  16,   8,  17,   4,   6,   9,   4,  28,   8,  53,  24,   2,
        31,  11,  10,  38,   3,  13,   3,  17,  80,   7,   2,  19,  22,
         5,  16,   2,  13,   2,  19,   7,  95,  14,   8,   2,   1,  38,
        19,   3,   5,  27,  19,   8,   8,  10,  11,   3,  13,  16,  25,
        13,   7,  17,   4,   6,   7,   5,  20,   7,  47,  23,   2,  28,
        10,  10,  32,   2,  11,   2,  16,  71,   7,   2,  20,  18,   5,
        17,   2,  11,   1,  18,   6,  85,  14,   8,   2,   1,  39,  20,
         2,   5,  23,  16,   6,   7,   8,  10,   3,  12,  15,  19,  10,
         6,  13,   3,   4,   7,   4,  23,   4,  39,  20,   2,  25,   7,
        10,  25,   2,   9,   2,  13,  66,   7,   2,  17,  15,   4,  17,
         2,   8,   2,  16,   7,  83,  13,   8,   2,   1,  38,  18,   2,
         4,  25,  14,   6,   6,   8,   9,   2,  12,  15,  21,   9,   5,
        11,   2,   4,   7,   3,  23,   4,  36,  16,   2,  22,   7,   9,
        24,   2,   9,   2,  11,  70,   7,   1,  14,  15,   4,  14,   1,
         6,   2,  13,   5,  71,  11,   8,   1,   1,  31,  15,   2,   4,
        23,  11,   6,   6,   7,  10,   2,   9,  12,  18,   9,   4,  10,
         2,   4,   6,   2,  15,   5,  31,  13,   1,  20,   7,   8,  20,
         2,   8,   2,   9,  57,   6,   1,  13,  12,   3,  12,   1,   6,
         1,  10,   5,  61,   9,   7,   1,   1,  29,  11,   2,   3,  24,
         9,   5,   6,   5,   8,   2,   7,  11,  14,   7,   3,   9,   3,
         3,   5,   2,  11,   4,  25,  10,   1,  15,   6,   9,  14,   2,
         6,   2,   8,  39,   5,   1,  12,  12,   2,  10,   1,   6,   1,
        10,   5,  56,  10,   5,   1,   1,  22,  12,   2,   3,  19,   8,
         5,   5,   5,   7,   2,   6,  10,  12,   7,   3,   8,   2,   3,
         4,   2,  10,   3,  20,  12,   1,  15,   5,   7,  15,   1,   6,
         2,   9,  32,   5,   1,  10,  13,   2,   9,   1,   6,   1,  10,
         4,  54,   9,   5,   1,   0,  18,  12,   2,   3,  15,   9,   5,
         5,   6,   5,   2,   6,   8,  12,   7,   3,   8,   2,   3,   4,
         2,  10,   3,  22,  12,   1,  14,   5,   7,  15,   1,   5,   1,
         8,  36,   5,   1,  10,  11,   2,   8,   1], dtype=int64)
```

In [55]:
```
#total outlier analysis of given data

len(Total)
```

Out[55]:
```
867
```

In [56]:
```
#cluster analysis

Total_column2 = df.loc[:,'Totals.Alcohol.Use Disorder Past Year.18-25']
Total1 = Total_column.values
Total1
```

Out[56]:
```
array([ 18,   4,  36,  14, 173,  26,  16,   4,   1,  73,  36,   7,  10,
        66,  34,  19,  12,  17,  24,   6,  26,  37,  57,  32,  10,  31,
         9,  14,  13,   9,  37,  14,  87,  39,   6,  57,  17,  17,  55,
         6,  17,   7,  23, 118,  11,   3,  38,  31,   9,  35,   3,  18,
         4,  37,  13, 204,  26,  19,   3,   1,  77,  32,   7,   9,  66,
        31,  18,  17,  18,  24,   7,  25,  34,  54,  30,  10,  35,   8,
        12,  13,   7,  41,  14,  86,  32,   5,  55,  20,  19,  58,   6,
        19,   7,  22, 113,  11,   3,  33,  34,   9,  42,   3,  21,   4,
        36,  12, 200,  29,  19,   3,   1,  81,  33,   5,   9,  64,  33,
        18,  18,  20,  20,   7,  25,  27,  51,  27,  11,  34,   7,  11,
        13,   7,  42,  12,  85,  38,   4,  53,  18,  19,  58,   5,  17,
         5,  23, 108,  12,   3,  31,  30,   9,  38,   3,  19,   4,  30,
        15, 182,  27,  18,   3,   1,  73,  36,   5,   9,  62,  28,  20,
        17,  20,  17,   6,  25,  28,  47,  29,  11,  31,   8,  11,  13,
         7,  40,  11,  84,  35,   4,  52,  15,  19,  53,   5,  16,   5,
        24, 104,  12,   3,  31,  31,   7,  29,   3,  19,   4,  26,  17,
       186,  28,  21,   3,   1,  69,  35,   6,   9,  64,  27,  18,  16,
        17,  17,   6,  25,  30,  50,  31,  10,  29,   6,  10,  13,   7,
        38,  11,  74,  31,   4,  55,  14,  16,  50,   5,  18,   5,  27,
       105,  11,   3,  30,  30,   6,  30,   3,  17,   3,  32,  12, 179,
        27,  18,   3,   1,  66,  32,   5,   8,  56,  28,  15,  14,  14,
        15,   5,  22,  27,  47,  29,  10,  28,   5,   9,  13,   6,  32,
        10,  74,  31,   3,  53,  14,  17,  48,   4,  16,   4,  22, 103,
        10,   3,  32,  29,   6,  30,   3,  16,   3,  29,  11, 166,  24,
        17,   3,   1,  58,  28,   5,   7,  46,  27,  16,  13,  13,  13,
         4,  17,  24,  41,  24,   9,  25,   6,   8,  12,   5,  26,  11,
        82,  25,   3,  49,  15,  18,  41,   4,  14,   4,  16,  98,   8,
         2,  31,  25,   6,  24,   3,  16,   3,  26,  11, 167,  22,  14,
         3,   1,  59,  30,   5,   7,  43,  23,  13,  12,  14,  12,   4,
        15,  25,  37,  20,   9,  22,   5,   7,  11,   5,  27,   9,  75,
        26,   2,  43,  14,  15,  40,   4,  14,   4,  18,  97,  10,   2,
        30,  22,   6,  20,   2,  14,   3,  25,   9, 172,  18,  13,   3,
         1,  55,  31,   5,   6,  41,  19,  10,  10,  12,  12,   3,  15,
        24,  33,  18,   9,  20,   4,   5,  11,   5,  33,   7,  62,  29,
         2,  36,  11,  12,  40,   3,  14,   3,  18,  85,  10,   2,  26,
        24,   5,  17,   2,  13,   2,  20,   7, 133,  15,  11,   2,   1,
        42,  25,   4,   5,  37,  19,   8,   8,  11,  12,   3,  15,  20,
        31,  16,   8,  17,   4,   6,   9,   4,  28,   8,  53,  24,   2,
        31,  11,  10,  38,   3,  13,   3,  17,  80,   7,   2,  19,  22,
         5,  16,   2,  13,   2,  19,   7,  95,  14,   8,   2,   1,  38,
        19,   3,   5,  27,  19,   8,   8,  10,  11,   3,  13,  16,  25,
        13,   7,  17,   4,   6,   7,   5,  20,   7,  47,  23,   2,  28,
        10,  10,  32,   2,  11,   2,  16,  71,   7,   2,  20,  18,   5,
        17,   2,  11,   1,  18,   6,  85,  14,   8,   2,   1,  39,  20,
         2,   5,  23,  16,   6,   7,   8,  10,   3,  12,  15,  19,  10,
         6,  13,   3,   4,   7,   4,  23,   4,  39,  20,   2,  25,   7,
        10,  25,   2,   9,   2,  13,  66,   7,   2,  17,  15,   4,  17,
         2,   8,   2,  16,   7,  83,  13,   8,   2,   1,  38,  18,   2,
         4,  25,  14,   6,   6,   8,   9,   2,  12,  15,  21,   9,   5,
        11,   2,   4,   7,   3,  23,   4,  36,  16,   2,  22,   7,   9,
        24,   2,   9,   2,  11,  70,   7,   1,  14,  15,   4,  14,   1,
         6,   2,  13,   5,  71,  11,   8,   1,   1,  31,  15,   2,   4,
        23,  11,   6,   6,   7,  10,   2,   9,  12,  18,   9,   4,  10,
         2,   4,   6,   2,  15,   5,  31,  13,   1,  20,   7,   8,  20,
         2,   8,   2,   9,  57,   6,   1,  13,  12,   3,  12,   1,   6,
         1,  10,   5,  61,   9,   7,   1,   1,  29,  11,   2,   3,  24,
         9,   5,   6,   5,   8,   2,   7,  11,  14,   7,   3,   9,   3,
         3,   5,   2,  11,   4,  25,  10,   1,  15,   6,   9,  14,   2,
         6,   2,   8,  39,   5,   1,  12,  12,   2,  10,   1,   6,   1,
        10,   5,  56,  10,   5,   1,   1,  22,  12,   2,   3,  19,   8,
         5,   5,   5,   7,   2,   6,  10,  12,   7,   3,   8,   2,   3,
         4,   2,  10,   3,  20,  12,   1,  15,   5,   7,  15,   1,   6,
         2,   9,  32,   5,   1,  10,  13,   2,   9,   1,   6,   1,  10,
         4,  54,   9,   5,   1,   0,  18,  12,   2,   3,  15,   9,   5,
         5,   6,   5,   2,   6,   8,  12,   7,   3,   8,   2,   3,   4,
         2,  10,   3,  22,  12,   1,  14,   5,   7,  15,   1,   5,   1,
         8,  36,   5,   1,  10,  11,   2,   8,   1], dtype=int64)
```

In [57]:
```
len(Total1)
```

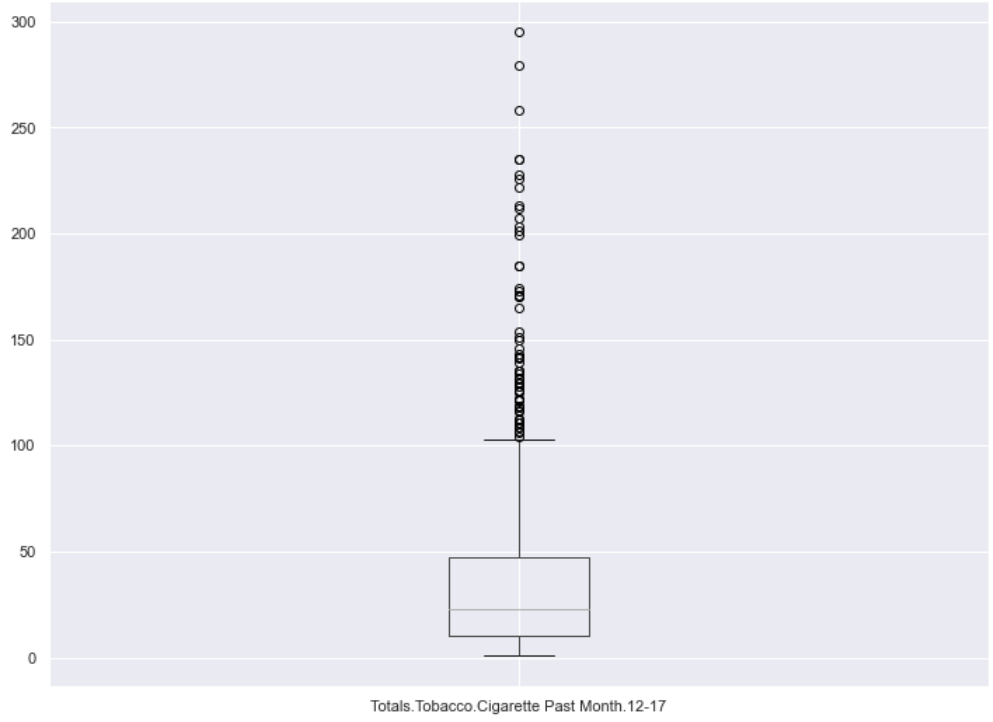
Out[57]:
```
867
```

In [58]:
```
#Using Z score

outliers=[]
def detect_outlier(Total_1):

    threshold=3
    mean_1 = np.mean(Total_1)
    std_1 =np.std(Total_1)


    for y in Total_1:
        z_score= (y - mean_1)/std_1
        if np.abs(z_score) > threshold:
            outliers.append(y)
    return outliers
```

In [59]:
```
outlier_datapoints = detect_outlier(Total)
len(outlier_datapoints)
```

Out[59]: 18

In [60]:
```
#using IQR

df.boxplot(column=['Totals.Tobacco.Cigarette Past Month.12-17'])
x = plt.show
```



In [61]:
```
#using IQR

sorted(Total)
```

Out[61]:
```
[0,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
 1,
```

In [62]:
```
#USING IQR

q1, q3= np.percentile(Total,[25,75])
print(q1,q3)
```

```
5.0 24.0
```

In [63]:
```
#USING IQR

iqr = q3-q1
iqr
```

Out[63]:
```
19.0
```

In [64]:
```
#USING IQR

lower_bound = q1 - (1.5 * iqr)
lower_bound
```

Out[64]:
```
-23.5
```

In [65]:
```
#USING IQR

upper_bound = q3 + (1.5 * iqr)
upper_bound
```

Out[65]:
```
52.5
```

In [66]:
```
df['Totals.Marijuana.New Users.18-25'] = np.where(df['Totals.Marijuana.New Users.18-25']>upper_bound,upper_bound,df['T
df['Totals.Marijuana.New Users.18-25'] = np.where(df['Totals.Marijuana.New Users.18-25']<lower_bound,lower_bound,df['T
```

In [67]:
```
df.boxplot(column=['Totals.Marijuana.New Users.18-25'])
x = plt.show
```

In [68]:
```
#Linear regression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OrdinalEncoder
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2
from sklearn.feature_selection import mutual_info_classif
from matplotlib import pyplot
```

In [69]:
```
def prepare_inputs(X_train, X_test):
    oe = OrdinalEncoder()
    oe.fit(X_train)
    X_train_enc = oe.transform(X_train)
    X_test_enc = oe.transform(X_test)
    return X_train_enc, X_test_enc
```

In [70]:
```
def prepare_targets(y_train, y_test):
    le = LabelEncoder()
    le.fit(y_train)
    y_train_enc = le.fit_transform(y_train)
    y_test_enc = le.fit_transform(y_test)
    return y_train_enc, y_test_enc
```

In [71]:
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from sklearn.linear_model import LinearRegression
```

In [74]:
```
bihar_df = df[df['State'] == 'Alabama'].copy()
bihar_df = bihar_df.groupby(['Year']).sum()
bihar_df.reset_index(inplace=True)
```

In [75]:
```
#Building the model to predict the actual total number of cases in

X = bihar_df[['Year']]
y = bihar_df['Totals.Alcohol.Use Disorder Past Year.18-25']

#Splitting the dataset
X_train,X_test,Y_train,Y_test=train_test_split(X,y,test_size=.3,random_state=10)

lin=LinearRegression()
lin.fit(X_train,Y_train)
ypred=lin.predict(X_test)

df = pd.DataFrame({'Actual Value': Y_test, 'Predicted Value': ypred})
df
```

Out[75]:

	Actual Value	Predicted Value
5	65	62.786146
3	66	65.820231
14	41	49.132765
7	71	59.752062
6	70	61.269104
8	64	58.235019

In [76]:
```
X = bihar_df[['Year']]
y = bihar_df['Totals.Tobacco.Cigarette Past Month.12-17']

from sklearn.model_selection import train_test_split
X_train,X_test,Y_train,Y_test=train_test_split(X,y,test_size=.25,random_state=10)

model = LinearRegression()
model.fit(X, y)

data = [[2001]]
X_predict = pd.DataFrame(data, columns = ['Year'])

y_predict = model.predict(X_predict)

df = pd.DataFrame({'Predicted Value': y_predict})
df
```

Out[76]:

	Predicted Value
0	55.889706

In [ ]:
```
```

Key Insights

- By considering above research on this topic we came to know that drugs and alcohol usage rates are increasing gradually.
- On average the drug usage rate is increased with age and most of the young age people are addicted to it.
- Of those countries above mentioned that shows clearly the rate of usage of drugs in some particular countries are increasing more.

- This means in some states like California, Alabama, Texas etc., the drugs and alcohol usage among the people is increasing more due to their perspective reasons as we discussed above.
- Comparing with other drugs cocaine usage is less.
- From the above analysis we can conclude that Alcohol usage is decreasing gradually and parallelly drugs like illicit, marijuana, tobacco drugs usage is increasing.
- We have predicted future usage of alcohol and drugs usage.
- In above codes we have been predicted usage of 18-25,12-17 year age groups for alcohol and tobacco usage.From the prediction we noticed that in future maximum usage of drugs is slightly increasing
- some states like Wyoming and district of Columbia etc., the drug usage rate is low when compared to other states.

CONCLUSION

Our project consists of extensive information visualization to perform investigations on the data so as to discover certain patterns, to identify anomalies, and to check assumptions with the help of summary statistics, graphical representations and other valuable insights. The characteristics of the data have been summarized and the trends have been plotted. The results obtained give us a clear vision about what type of population is highly affected by this drug addiction. In the prediction part it is evident that, if some steps are not taken, the number of users will continue to be high, and we will end up losing students' lives in this illicit drug addiction. This analysis gives us a deeper apprehension about drug usage among the society in U.S. By understanding and evaluating the problems and situations to the students by performing some anti – drug addiction campaign they come to know how dangerous it is. And through rehabilitation centre's we can cure them. Future research among both male and female substance users at community level, including health professionals' experiences, is recommended to highly using countries.

Data set link:

https://corgis-edu.github.io/corgis/csv/drugs/ (https://corgis-edu.github.io/corgis/csv/drugs/)

References:

https://www.sciencedirect.com/science/article/abs/pii/S0733862720302534
(https://www.sciencedirect.com/science/article/abs/pii/S0733862720302534)

https://www.medindia.net/patientinfo/impact-of-drug-abuse-on-health-and-society.htm#tips-on-how-to-deal-with-drug-abuse
(https://www.medindia.net/patientinfo/impact-of-drug-abuse-on-health-and-society.htm#tips-on-how-to-deal-with-drug-abuse)

https://www.healthinaging.org/a-z-topic/drug-and-substance-use (https://www.healthinaging.org/a-z-topic/drug-and-substance-use)

In [ ]: